Received: 2 April 2022 | Revised: 31 October 2022 | Accepted: 12 December 2022

DOI: 10.1111/sjos.12632

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

Variable selection for high-dimensional generalized linear model with block-missing data

Yifan He[1] | **Yang Feng**[2] | **Xinyuan Song**[1]

[1]Department of Statistics, The Chinese University of Hong Kong, Hong Kong

[2]School of Global Public Health, New York University, New York, New York, USA

Correspondence
Xinyuan Song, Department of Statistics, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong
Email: xysong@sta.cuhk.edu.hk

Funding information
National Natural Science Foundation of China, Grant/Award Number: 11871263; Research Grants Council, University Grants Committee, Grant/Award Numbers: 14301918, 14302519

Abstract

In modern scientific research, multiblock missing data emerges with synthesizing information across multiple studies. However, existing imputation methods for handling block-wise missing data either focus on the single-block missing pattern or heavily rely on the model structure. In this study, we propose a single regression-based imputation algorithm for multiblock missing data. First, we conduct a sparse precision matrix estimation based on the structure of block-wise missing data. Second, we impute the missing blocks with their means conditional on the observed blocks. Theoretical results about variable selection and estimation consistency are established in the context of a generalized linear model. Moreover, simulation studies show that compared with existing methods, the proposed imputation procedure is robust to various missing mechanisms because of the good properties of regression imputation. An application to Alzheimer's Disease Neuroimaging Initiative data also confirms the superiority of our proposed method.

KEYWORDS

block-wise missingness, graph model, inverse covariance matrix, Lasso, sparsity

1 | INTRODUCTION

With widespread data transmission and integration, multimodality or multisource data emerge from modern scientific research. Epidemiological analysis, clinical trials, and genome-wide studies all need to synthesize complementary information across multiple type testings. However, a common characteristic of multimodality data is that groups of observation can be missing entirely for a specific modality. In the study on Alzheimer's Disease (AD) Neuroimaging Initiative (ADNI) study, almost all subjects have magnetic resonance imaging (MRI) information to detect and diagnose the progression of AD. However, not all of the subjects are willing to consent to an invasive procedure. For instance, only half of the subjects are collected for cerebrospinal fluid (CSF); many lack proteomics information, especially with normal cognition. Such kind of block-wise missing pattern hampers the analysis of multimodality data.

The most straightforward approach is to pool observations from multiple sources together and only keep the complete data. However, discarding subjects with missing measures may challenge underlying randomization rules or lose a great deal of information, let alone deal with the situation that complete cases are very few or do not exist. Therefore, several methods have been recently developed to fully take advantage of multimodality data and target block-wise missing settings. The first class of approaches was inspired by multitask learning (Ando et al., 2005; Argyriou et al., 2008; Liu et al., 2012). Yuan et al. (2012) proposed the incomplete multisource feature (iMSF) learning method. Taking the ADNI data above as example, if all samples accept MRI measurement, then subjects can be classified into four groups based on their missing patterns: (1) CSF, MRI; (2) proteome, MRI; (3) CSF, proteome, MRI; and (4) MRI. iMSF transforms multiple-modality learning into four prediction tasks with four different regression models and combines them by minimizing the summation of the loss functions. One strong constraint involved in iMSF is that regression models, including a specific modality, shares the same set of important predictors in this modality. This assumption would be inappropriate when modalities are highly correlated. Another challenge is that the model dimension involved in iMSF grows with the number of modalities exponentially. To address this issue, Xiang et al. (2014) and Li et al. (2018) refined and recombined the subtasks by utilizing shared covariates repeatedly and imposing different weights for their corresponding parameters. As a result, the number of parameters decreased through a superposed weight vector. However, explaining the bilevel coefficients in real-world applications is difficult, and the coefficients of the whole model (CSF, proteome, and MRI) are inaccessible if no complete case exists. Moreover, other existing multimodality data learning methods heavily rely on the structure of the linear regression model. For example, Yu et al. (2020) proposed a direct optimization approach that needed to estimate the covariance matrix of predictors and the cross-covariance vector between the predictors and the responses. Inspired by the single regression imputation method, Xue and Qu (2021) and Xue et al. (2021) permutated the original covariate matrices into multiple one-block missing submatrices for the imputation step. Consequently, unbiased estimating equations or other combination approaches for deriving overlapped imputation values are crucial in this processing mode. Thus, their method is model-based and not easy to generalize, and a highly flexible prediction method for block-missing multimodality data should be developed.

This study proposes a two-step algorithm to analyze block-missing multimodality data under a generalized linear model (GLM). In the first step, we estimate covariance matrix or precision matrix under high-dimensional multivariate normality assumption with block-wise missing data,

thus allowing for imputation with mean conditional on observed blocks. Several methods are available to estimate the covariance matrix. For complete data Friedman et al. (2008), Yuan and Lin (2007), and Banerjee et al. (2008) proposed a group least absolute shrinkage and selection operator (GLasso) approach, which provided ℓ_1-norm regularized maximum-likelihood estimation of precision matrix. Lam and Fan (2009) and Fan et al. (2009) further developed extended versions: SCAD GLasso and adaptive GLasso. For missing data, Städler and Bühlmann (2012) proposed MissGLasso estimator of the precision matrix when only one set of predictors was missing. Given that all the above methods need a sample covariance matrix as an initial estimator, Yu et al. (2020) reconstructed the sample covariance matrix as a linear combination of the identity matrix, the estimates of the intra-modality covariance matrix, and the cross-modality covariance matrix for multimodality data. However, the induced coefficients complicate the calculation. Therefore, we take the same spirit of Yu et al. (2020)'s work but simplify the procedure to estimate the covariance matrix and examine its sparsity via the Lasso-type penalized likelihood. In the second step, we impose a sparsity penalty to estimate the coefficients for the high-dimensional GLM based on the imputed data.

This study contributes to the existing literature in three aspects. First, we design an efficient imputation algorithm to address multiple block-wise missing data that may not contain any complete case. When complete cases existing, our method can be seen as an improvement of traditional regression imputation, since the information of incomplete groups is encoded in our final estimation. A sparse covariance matrix estimate is a byproduct of the proposed algorithm. To the best of our knowledge, this study is the first one to provide sparse covariance matrix estimation tailored to the multiblock missing data. Second, the proposed method conducts prediction for high-dimensional GLM with multiblock missing data. In contrast, most existing works (e.g., Xiang et al., 2014; Xue & Qu, 2021; Yu et al., 2020) have mainly focused on linear model prediction, thereby becoming invalid in the presence of a binary or categorical response. An exception in this direction is the work of Yu and Hou (2022), who proposed a weighted nearest neighbor classifier to solve the binary classification problem. However, their method requires the existence of complete cases. Moreover, the imputation step in the proposed method is independent of the subsequent prediction task. Therefore, our approach can be easily extended to various circumstances with different prediction targets. Finally, we establish theoretical results for the proposed estimator in the GLM context. In particular, the convergence rate of our estimator is comparable with that derived in Fan and Lv (2013) and Xue and Qu (2021).

The remainder of this paper is organized as follows. Section 2 introduces the motivation and the proposed method. Section 3 presents the theoretical results about the estimates of the covariance matrix and establishes the predictor selection consistency. Sections 4 and 5 demonstrate the performance of the proposed method through simulation studies and an application to the ADNI dataset, respectively. Section 6 concludes the paper. All technical proofs are provided in Appendix S1.

2 | MOTIVATION AND METHODOLOGY

Some generic notations used in this paper are collected here. If A is a $n \times m$ matrix, then $A_{B,D}$ denotes a submatrix of A, where B and D are some sets of row and column indices, respectively. For vector a, a_b denotes a subvector of a, where b is an index set.

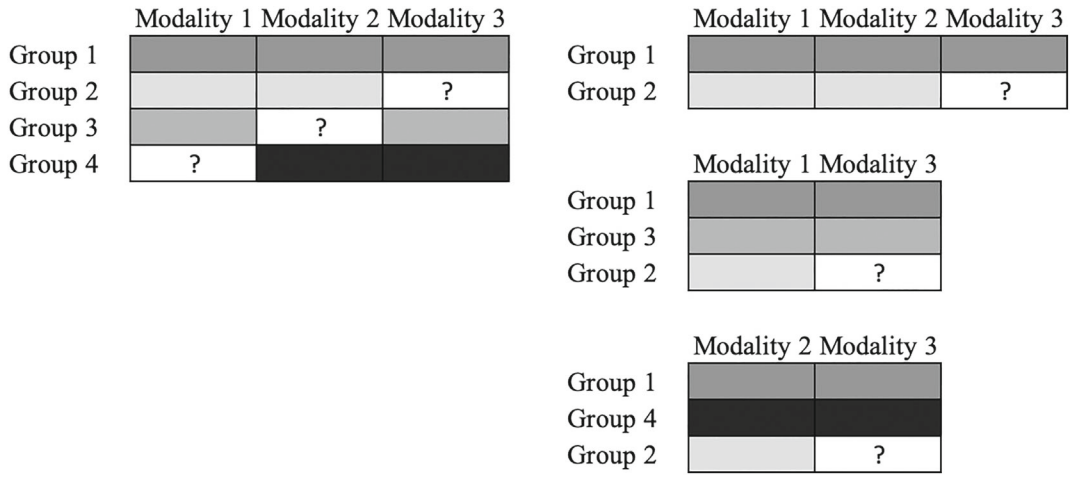

FIGURE 1 Left panel: an example of block-missing multimodality data. Right panel: single-block missing pattern. The shaded blocks are observed while the white blocks are missing.

2.1 | Background and problem setup

Let $Y = (y_1,\ldots,y_n)^\top$ be a sample of response vector y, and $X = (X_1,\ldots,X_p)$ be an $n \times p$ complete design matrix, of which rows are independent samples of a random vector $x = (x_1,\ldots,x_p)$. The covariance matrix of x, denoted as Σ, is positive definite. On the basis of the missing patterns across all modalities, we divide samples into R disjoint groups. Let $n(r)$ denote the number of observations allocated into the rth group, and $n = \sum_{r=1}^{R} n(r)$. For $r = 1,\ldots,R$, let $p(r)$ be the number of completely observed predictors in the rth group, $G(r)$ be the index set of observations in the rth group, and $X_{G(r),o(r)}$ and $X_{G(r),m(r)}$ be the observed and missing parts of the design matrix in the rth group, where $o(r)$ and $m(r)$ are the index sets of the observed and missing covariates, respectively. They are marked as shaded blocks and white blocks in Figure 1. For simplicity, we omit subscript $G(r)$ and abbreviate $X_{G(r),o(r)}$ and $X_{G(r),m(r)}$ as $X_{o(r)}$ and $X_{m(r)}$, respectively, when the context is clear. Let $X_{G(r)} = X_{G(r),m(r)\cup o(r)}$ denote the samples in Group r.

The GLM assumes that the conditional distribution of y given x belongs to the exponential family with the following density function:

$$f(y;\eta,\phi) = \exp\{y\eta - b(\eta) + c(y,\phi)\}, \tag{1}$$

where $\eta = x^\top\beta$, $\beta = (\beta_1,\ldots,\beta_p)^\top \in \mathbb{R}^p$ is the regression coefficient vector, $b(\cdot)$ and $c(\cdot,\cdot)$ are known functions, and $\phi > 0$ is the dispersion parameter. Assume that $b(\cdot)$ is smooth and convex with $b''(\cdot)$ is bounded away from 0 and ∞. Up to an affine transformation, the log-likelihood function given by the sample is

$$\ell_n(\beta) = n^{-1}[Y^\top X\beta - \mathbf{1}^\top b(X\beta)].$$

To ensure identifiability and interpretability for a high-dimensional model, we typically assume that the true regression coefficient vector β^* is sparse and consists of only s_β nonzero components. Let $S_\beta = \{j : \beta_j^* \neq 0\}$ and $S_\beta^c = \{j : \beta_j^* = 0\}$ correspond to relevant and irrelevant covariates, respectively.

Under the GLM setting, coefficients in β can be estimated by minimizing the penalized negative log-likelihood function as follows:

$$Q_n(\beta) = -n^{-1}[Y^\top X\beta - \mathbf{1}^\top b(X\beta)] + \sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{2}$$

where $p_\lambda(t)$ with $t \in [0, \infty)$ is a penalty function indexed by the regularization parameter λ. To ease the presentation, λ in (2) is denoted as λ_β. Here, we select the Lasso penalty as an example. Considering that the focus of our method is the imputation step, we can easily apply our method to other regularization forms.

2.2 | Regression imputation with single-block missing pattern

Regression imputation is a common approach for handling missing covariates. It substitutes $X_{m(r)}$, $r = 1,\ldots,R$ by its mean conditional on the observed covariates, $E(X_{m(r)}|X_{o'(r)})$, where $o'(r) \subset o(r)$. The estimate of $E(X_{m(r)}|X_{o'(r)})$, denoted as $\widehat{E}(X_{m(r)}|X_{o'(r)})$, heavily relies on a kind of single-block missing (SBM) structure, as shown in the right panel of Figure 1, regardless of specific estimation methods. In the literature, some low-rank matrix completion methods (Cai et al., 2016; Mazumder et al., 2010) relied on an SBM to recover missing entries. Then, a question about how to choose $o'(r)$ or SBM arises immediately. One traditional approach is to take $o'(r) = o(r)$ and only use complete observations to estimate the cross correlation between $x_{m(r)}$ and $x_{o(r)}$ (e.g., Städler & Bühlmann, 2012). As shown in the left panel of Figure 1, Groups 1 and 2 form a SBM, which can be utilized to estimate $E(X_{m(2)}|X_{o(2)})$.

However, this traditional approach is applicable only when enough complete observations exist. Moreover, the information incorporated in other groups is omitted to some extent in the imputation step. Xue and Qu (2021) adopted a different approach and selected all available $o'(r)$, where $o'(r)$ is not unique. In other words, by rearranging the original data, they utilized all SBMs for imputation. For example, as shown in Figure 1 (right panel), three SBMs are used for imputing $X_{m(2)}$. This method results in several problems. First, the whole imputation efficiency may be dramatically affected by the minimum sample size group. Moreover, the number of SBMs could grow exponentially with the number of modalities, thereby increasing the computation complexity. Finally, many homogeneous imputation values generated by the above process need to be integrated. Directly stacking all the values increases the computation burden and leads to a biased estimator because it destroys the randomness of the experiment design.

2.3 | Regression imputation with multi-block missing pattern

Suppose that x follows a joint Gaussian distribution with mean τ and covariance matrix Σ. This distribution assumption is used only to motivate our idea. Later, we will show that the implementation is entirely independent of this assumption. The precision matrix of x is denoted by Θ. Then, given $x_{o(r)}$, $r = 1,\ldots,R$, $x_{m(r)}$ follows a Gaussian distribution with mean $\tau_{m(r)} + \Sigma_{m(r),o(r)}\Sigma_{o(r),o(r)}(x_{o(r)} - \tau_{o(r)})$ and covariance $\Sigma_{m(r),m(r)} - \Sigma_{m(r),o(r)}\Sigma^{-1}_{o(r),o(r)}\Sigma_{o(r),m(r)}$ (Lauritzen, 1996). Thus, in terms of Θ, the conditional distribution can be re-expressed by (Städler & Bühlmann, 2012)

$$x_{m(r)}|x_{o(r)} \sim \mathcal{N}\left(\tau_{m(r)} - \Theta^{-1}_{m(r),m(r)}\Theta_{m(r),o(r)}(x_{o(r)} - \tau_{o(r)}), \Theta^{-1}_{m(r),m(r)}\right).$$

For the missing mechanism, we take a setting similar to that in Yu et al. (2020). We define $O_{ij} = \mathbf{1}\{X_{ij} \text{ is observed}\}$, and $N_{\min} = \min_{j,k\in\{1,\ldots,p\}} \sum_{i=1}^{n} O_{ij}O_{ik}$. Assume that the first sample moment

$$\hat{\tau}_j = \sum_{i=1}^{n} O_{ij}X_{ij} / \sum_{i=1}^{n} O_{ij}, \quad j = 1,\ldots,p,$$

and the second sample moment

$$\hat{\Sigma}_{jk} = \sum_{i=1}^{n} O_{ij}X_{ij}O_{ik}X_{ik} / \sum_{i=1}^{n} O_{ij}O_{ik}, \quad j,k \in \{1,\ldots,p\},$$

are unbiased estimators of $E(x_j)$ and $E(x_jx_k)$, respectively.

Without loss of generality, let $\tau = 0$. That is, the observations of every predictor are centered to have mean 0. Then, the best linear unbiased estimate of $X_{m(r)}$ is $-X_{o(r)}\Theta_{o(r),m(r)}\Theta^{-1}_{m(r),m(r)}$. Thus, deriving the estimate of the precision matrix with multiblock missing data is desirable.

2.4 | Sparse estimate of Θ

Suppose that each column of Θ is k-sparse. Then, to bound the error of the estimator of Θ in ℓ_1, our method bears similarity to the approach of Yuan (2010) but is valid in the case of multiblock incomplete data. Let X_{ij} denote the element of X at the ith row and jth column, and $X_{i,-j}$ denote the ith row of X with the jth column removed. Under the high-dimensional Gaussian graphical model setting, a vector $\theta^{(j)} \in \mathbb{R}^{p-1}$ exists, such that

$$X_{ij} = X_{i,-j}\theta^{(j)} + \varepsilon_{ij}, \quad i = 1,\ldots,n,\ j = 1,\ldots,p,$$

where ε_{ij} is independent of $X_{i,-j}$ and $\mathrm{Var}(\varepsilon_{ij}) = \Theta^{-1}_{jj}$. Let $\theta^{(j)}_k$ be the kth element of $\theta^{(j)}$. Since $\theta^{(j)}_k = \Theta_{jk}/\Theta_{kk}$, then we have

$$\Theta_{jj} = (\mathrm{Var}(\varepsilon_{ij}))^{-1}, \quad \Theta_{-j,j} = -(\mathrm{Var}(\varepsilon_{ij}))^{-1}\theta^{(j)}, \tag{3}$$

where $\Theta_{-j,j}$ is the jth column of Θ with the jth row removed. Thus, a sparse estimator of Θ can be obtained by regressing X_{ij} over $X_{i,-j}$, $i = 1,\ldots,n$ with a neighborhood selection (Meinshausen & Bühlmann, 2006), if the data are complete. Specifically, the Lasso estimate of $\theta^{(j)}$ is given by

$$\arg\min_{\theta^{(j)}\in\mathbb{R}^{p-1}} \left\{ n^{-1}\sum_{i=1}^{n}(X_{ij} - X_{i,-j}\theta^{(j)})^2 + \lambda_\theta\|\theta^{(j)}\|_1 \right\}. \tag{4}$$

To apply this estimation procedure to multiblock missing data, following Witten and Tibshirani (2009), we reformulate the optimization problem in (4) as

$$\hat{\theta}^{(j)} = \arg\min_{\theta^{(j)}\in\mathbb{R}^{p-1}} \left\{ \frac{1}{2}\theta^{(j)\top}\tilde{\Sigma}_{-j,-j}\theta^{(j)} - \tilde{\Sigma}^{\top}_{-j,j}\theta^{(j)} + \lambda_\theta\|\theta^{(j)}\|_1 \right\}, \tag{5}$$

where $\tilde{\Sigma}$ is an initial estimator of Σ, $\tilde{\Sigma}_{-j,j}$ is defined in the same manner as $\Theta_{-j,j}$ in (3), and $\tilde{\Sigma}_{-j,-j}$ is the submatrix of $\tilde{\Sigma}$ with the jth row and jth column removed. We assume that $\tilde{\Sigma}$ satisfies the following condition:

Condition 1. ("Good" approximation) $\Pr\left(\|\tilde{\Sigma}-\Sigma\|_{\max} \geqslant V_1\sqrt{\log p/N_{\min}}\right) \leqslant V_2(n,p)$, $V_1 > 0$ and $V_2(n,p) \to 0$.

If x is sub-Gaussian, then we can take the sample covariance matrix as the initial estimator, that is, $\tilde{\Sigma} = (\hat{\Sigma}_{jk})$, $j,k = 1,\ldots,p$. If x follows a heavy-tailed distribution, then robust initial estimators, such as the median of mean estimator

$$\tilde{\Sigma} = (\tilde{\sigma}_{jk}), \text{ with } \tilde{\sigma}_{jk} = \text{median of } \left\{X_{ij}X_{ik},\ X_{ij} \text{ and } X_{ik} \text{ are observed}\right\}_{i=1}^{n},$$

can be adopted. Other robust initial estimators that satisfy Condition 1 can be found in Avella-Medina et al. (2018).

The optimization objective (5) is not convex when $\tilde{\Sigma}$ is not positive semi-definite due to missing blocks. Thus, according to Datta and Zou (2017), the nearest positive semi-definite matrix of $\tilde{\Sigma}$, given by

$$\hat{\Sigma}_{+} =: \arg\min_{\Sigma\in\Sigma_{pos}^{p}} \|\tilde{\Sigma}-\Sigma\|_{\max},$$

is a more suitable substitution, where Σ_{pos}^{p} is the set of positive semi-definite matrices. Finally, after replacing $\tilde{\Sigma}$ with $\hat{\Sigma}_{+}$, the estimate of $\theta^{(j)}$ can be obtained from (5) using the coordinate descent algorithm of Witten and Tibshirani (2009). In this study, we offer another efficient and convenient procedure.

We generate a random $B \times p$ matrix Z with entries being independently distributed according to the standard normal distribution, where $B = O(n)$ and $B > n$. In practice, $2n$ is an appropriate value for B. Let $\tilde{Z} = ZC$, where the upper triangular matrix C is the Cholesky factor of $\hat{\Sigma}_{+}$, that is, $C^{\top}C = \hat{\Sigma}_{+}$. Finally, by replacing X with $\tilde{Z}$ in (4), we can obtain estimates $\hat{\theta}^{(j)}$, $j = 1,\ldots,p$ with the following reformulation of (5):

$$\arg\min_{\theta^{(j)}\in\mathbb{R}^{p-1}} \left\{ B^{-1}\sum_{i=1}^{B}(\tilde{Z}_{ij} - \tilde{Z}_{i,-j}\theta^{(j)})^2 + \lambda_\theta\|\theta^{(j)}\|_1 \right\}, \tag{6}$$

and the corresponding mean squared error of the residual is

$$\widehat{\mathrm{Var}}(\varepsilon_{ij}) = B^{-1}\sum_{i=1}^{B}(\tilde{Z}_{ij} - \tilde{Z}_{i,-j}\hat{\theta}^{(j)})^2.$$

Naturally, the estimator of Θ can be obtained as follows:

$$\tilde{\Theta}_{jj} = (\widehat{\mathrm{Var}}(\varepsilon_{ij}))^{-1}, \quad \tilde{\Theta}_{-j,j} = -(\widehat{\mathrm{Var}}(\varepsilon_{ij}))^{-1}\hat{\theta}^{(j)}. \tag{7}$$

It can be seen that $\tilde{Z}$ is only used for copying the correlation structure among predictors. Since $\tilde{Z}$ is complete, estimating Θ with $\tilde{Z}$ will be more convenient than with original incomplete X.

Algorithm 1. Variable selection with block-missing data

Input: Y, X.

1 Compute the sample covariance matrix $\tilde{\Sigma}$ for X;
2 Find the nearest positive semi-definite matrix $\widehat{\Sigma}_+$ of $\tilde{\Sigma}$;
3 Generate a random $B \times p$ matrix Z with standard normally distributed entries;
4 Perform Cholesky Decomposition to $\widehat{\Sigma}_+$ and let $\tilde{Z} = ZC$, where $C^\top C = \widehat{\Sigma}_+$;
5 Compute $\hat{\theta}^{(j)}$ and the corresponding mean squared error $\widehat{\mathrm{Var}}(\varepsilon_{ij})$ with Lasso regression (6), for $j = 1, \ldots, p$;
6 Obtain estimate $\tilde{\Theta}$ by formula (7) and make symmetrization to get $\widehat{\Theta}$;
7 Replace the missing blocks in X with $\widehat{X}_{m(r)}$ and get the new design matrix $\widehat{X}$;
8 Solve optimization problem (9) and get $\hat{\beta}$;

Output: $\hat{\beta}$.

Considering the symmetry of Θ, we set $\widehat{\Theta} = \arg\min_{\Theta \in \Theta^p_{sym}} \|\Theta - \tilde{\Theta}\|_1$, where Θ^p_{sym} is the set of symmetric matrices.

Then, we replace the missing blocks in X with the imputation values for $X_{m(r)}$ as

$$\widehat{X}_{m(r)} = -X_{o(r)}\widehat{\Theta}_{o(r),m(r)}\widehat{\Theta}^{-1}_{m(r),m(r)} \tag{8}$$

and denote the new design matrix by $\widehat{X}$. In the final step, we can obtain the estimate $\hat{\beta}$ by solving the following optimization problem

$$\min_{\beta} \left\{ -n^{-1}[Y^\top \widehat{X}\beta - \mathbf{1}^\top b(\widehat{X}\beta)] + \lambda_\beta \|\beta\|_1 \right\}. \tag{9}$$

For clarity, the computing algorithm and pseudo code are described below (Algorithm 1).

3 | THEORETICAL RESULTS

In this section, we provide the theoretical foundation of the proposed method. Beforehand, some additional notation is collected. Let $\mathcal{K}(\Sigma) = \Lambda_{\max}(\Sigma)/\Lambda_{\min}(\Sigma)$, which is the condition number of Σ. Denote the index set of nonzero units in $\theta^{(j)}$ by $S^{(j)}_\theta$. Suppose the columns of the precision matrix Θ are k-sparse, where $k = \max_{j=1,\ldots,p} s^{(j)}_\theta$ and $s^{(j)}_\theta$ is the cardinality of $S^{(j)}_\theta$. The important covariate submatrix with nonzero coefficient is X_{S_β}. For Group r, $r = 1,\ldots,R$, let $\overline{X}_{G(r)}$ be the samples in Group r but with missing values replaced by

$$\tilde{X}_{m(r)}(\Theta) = -X_{o(r)}\Theta_{o(r),m(r)}\Theta^{-1}_{m(r),m(r)}$$

and $\tilde{X}^*_{m(r)} = \tilde{X}_{m(r)}(\Theta^*) = E(X_{m(r)}|X_{o(r)})$, where Θ^* is denoted as the true value of Θ. Let $\tilde{X} = (\tilde{X}^\top_{G(1)},\ldots,\tilde{X}^\top_{G(R)})^\top$, and $\mathcal{D}^* = \{\Theta : \|\Theta - \Theta^*\|_2 \leqslant d^*_0 \Lambda^{-2}_{\min}(\Sigma)\mathcal{K}^3(\Sigma)\sqrt{\log p/N_{\min}}\}$ be a neighborhood of Θ^* for some constant d^*_0. Note that the neighborhood is asymptotically shrinking if $\Lambda^{-2}_{\min}(\Sigma)\mathcal{K}^3(\Sigma)\sqrt{\log p/N_{\min}} \to 0$. Define function $\mu(\eta) = b'(\eta)$ and $e = Y - \mu(X_{S_\beta}\beta_{S_\beta})$. Let $\mathcal{B}^* = \{\beta : \|\beta - \beta^*\|_\infty \leqslant d^*_1\sqrt{\log p/n}\}$ be a neighborhood of β^* for some constant d^*_1.

We require the following regularity conditions to obtain the consistency of $\hat{\beta}$.

Condition 2. (Restricted eigenvalue [RE] for Σ) The covariance matrix Σ satisfies

$$\min_{\delta\in\{u\in\mathbb{R}^{p-1}:\|u_{S_\theta^{(j)c}}\|_1\le 7\|u_{S_\theta^{(j)}}\|_1\}}\frac{\delta^\top\Sigma_{-j,-j}\delta}{\delta^\top\delta}\geqslant m_j\geqslant m_{\min}>0,\ j=1,\ldots,p.$$

Condition 2 is similar to the condition (A2) in Yu et al. (2020). Its transformations for the covariance estimation and missing data can also be found in Yuan (2010), Datta and Zou (2017), and Loh and Wainwright (2012). This condition is used to obtain the bounds of the error of the clean Lasso estimate.

Condition 3. (Bounded variance) The function $b(\eta)$ satisfies that $C_{b,\min}\leqslant b(\eta)\leqslant C_{b,\max}$ in its domain, where $C_{b,\min}\leqslant C_{b,\max}$ are some positive constants.

Condition 3 is the same as the condition 2 in Fan and Lv (2013). It is a mild condition and commonly assumed in the GLM setting. The variances of all responses are bounded away from zero and infinity under Condition 3.

Condition 4. (Ultrahigh dimensionality) The dimension of covariates holds that

(1) $\log(p)=O(N_{\min}^{\alpha^*})$ for some constant $\alpha^*\in(0,1)$;
(2) $\Lambda_{\min}^{-2}(\Sigma)k^3\mathcal{K}^3(\Sigma)/m_{\min}\sqrt{\log p/N_{\min}}\to 0$;
(3) $kV_1\sqrt{\log p/N_{\min}}\to 0$.

Condition 4 allows the dimensionality p to increase up to exponentially fast with the minimum pairwise sample size $N_{\min}$, which keeps pace with the efficiency of the covariance estimator. This rate can also be found in Yu et al. (2020), where the eigenvalues of Σ are bounded away from zero and infinity.

Condition 5. For $\Theta\in\mathcal{D}^*$, and $\beta\in\mathcal{B}^*$, $\tilde{X}$ satisfies

$$0<C_{\tilde{X},\min}\leqslant\Lambda_{\min}\left(\frac{1}{n}\tilde{X}_{S_\beta}^\top\tilde{X}_{S_\beta}\right)\leqslant\Lambda_{\max}\left(\frac{1}{n}\tilde{X}_{S_\beta}^\top\tilde{X}_{S_\beta}\right)\leqslant C_{\tilde{X},\max}<\infty,$$

and

$$\left\|\left[\frac{1}{n}\tilde{X}_{S_\beta^c}^\top H(\tilde{X}_{S_\beta}\beta_{S_\beta})\tilde{X}_{S_\beta}\right]\left[\frac{1}{n}\tilde{X}_{S_\beta}^\top H(\tilde{X}_{S_\beta}\beta_{S_\beta})\tilde{X}_{S_\beta}\right]^{-1}\right\|_\infty\leqslant\nu_n<\infty,$$

where $H(\cdot)=\mathrm{diag}\{b''(\cdot)\}$ is a diagonal matrix.

The first part of Condition 5 assumes the lower and upper bounds for the eigenvalues of $\tilde{X}_{S_\beta}$ and is analogous to the robust spark condition in Fan and Lv (2013) and Xue and Qu (2021). Compared with the condition 5 of Xue and Qu (2021), our Condition 5 only controls the relevant part of $\tilde{X}$ rather than entire $\tilde{X}$. The second part of Condition 5 mainly controls the correlation between true and noise covariates. When $b''(\cdot)\equiv 1$, that is, the case of Gaussian linear model, the second part of Condition 5 is similar to the irrepresentable condition (Zhao & Yu, 2006).

Condition 6. (Error tail distribution) For $\Theta\in\mathcal{D}^*$, events $\mathcal{E}_0=\{\|n^{-1}\tilde{X}^\top e\|_\infty\leqslant\lambda_\beta/2\}$ and $\mathcal{E}_1=\{\|n^{-1}\tilde{X}_{S_\beta}^\top e\|_\infty\leqslant C_{\tilde{X},e}\sqrt{(\log n)/n}\}$ satisfy $P(\mathcal{E}_0)=1-O(p^{-c_e})$ and $P(\mathcal{E}_1)=1-O(n^{-c_e})$ for some positive constant c_e that can be sufficiently large for large enough $C_{\tilde{X},e}$.

Condition 6 is similar to the error tail condition in Fan and Lv (2013) and Fan et al. (2019). This condition can be satisfied with bounded or light-tailed errors. Details can be found in the appendix A of Fan and Lv (2013).

Condition 7. For $\Theta \in \mathcal{D}^*$, it holds with probability at least $1 - O(n^{-c_{\tilde{x}}})$ that

$$\left\| \frac{1}{n} \tilde{X}_{S_\beta}^\top (\tilde{X}_{S_\beta} - X_{S_\beta}) \beta_{S_\beta}^* \right\|_\infty \leqslant C_{\tilde{X},X} \sqrt{(\log n)/n},$$

and

$$\sup_{\tilde{\eta}} \left\| \frac{1}{n} \tilde{X}_{S_\beta}^\top H(\tilde{\eta}) (\tilde{X}_{S_\beta} - X_{S_\beta}) \beta_{S_\beta}^* \right\|_\infty \leqslant \lambda_\beta / 4,$$

where $\tilde{\eta}$ lies between $\tilde{X}_{S_\beta} \beta_{S_\beta}^*$ and $X_{S_\beta} \beta_{S_\beta}^*$.

Condition 7 is similar to the condition 7 in Xue and Qu (2021). A brief discussion on this condition and the proofs of the following theorems are provided in Appendix S1.

Theorem 1. *(Estimation loss for $\widehat{\Theta}$) Under Conditions 1–4, if $\lambda_\theta = O(V_1 \sqrt{\log p / N_{\min}})$, it holds that*

$$\|\widehat{\Theta} - \Theta^*\|_2 \leqslant O_p \left\{ \Lambda_{\min}^{-1}(\Sigma) k^3 \mathcal{K}^3(\Sigma) / m_{\min} \sqrt{\log p / N_{\min}} \right\}.$$

Theorem 1 states that $\widehat{\Theta}$ maintains estimation consistency when the number of covariates grows exponentially. If we assume that the eigenvalues of Σ are bounded, then the convergence rate is the same as that in Yu et al. (2020).

Theorem 2. *(Sign consistency and oracle inequality for $\hat{\beta}$) Under Conditions 1–7, if $\lambda_\beta = O(\sqrt{\log p / n})$, and $\nu_n \leqslant C_{\tilde{X},e}/8$, with probability at least $1 - O(n^{-C_2})$, it holds simultaneously that*

(a) (Sign consistency) $\operatorname{sgn}(\hat{\beta}) = \operatorname{sgn}(\beta^*)$;

(b) (Estimation loss) $\|\hat{\beta} - \beta^*\|_2 \leqslant C_1 C_{b,\min}^{-1} C_{b,\max} C_{\tilde{X},\min}^{-1} \sqrt{s_\beta (\log p)/n}$, *where C_1 is some positive constant and $C_2 = \min\{c_e, c_{\tilde{x}}\}$.*

Theorem 2 shows the sparsity and consistency of $\hat{\beta}$. The convergence rate in Theorem 2 is consistent with that obtained in Fan and Lv (2013) and Xue and Qu (2021).

4 | SIMULATION STUDY

In this section, we conduct simulation studies based on three commonly used GLMs, including linear, logistic, and Poisson regression models. For each model, we compare the proposed method with existing methods: (1) Stacked: The imputation method proposed by Xue and Qu (2021) imputes missing blocks multiple times and stacks the imputed results together. This method replaces $(X_{o(r)}, X_{m(r)})$ with $\left\{ \left(X_{o(r)}, \widehat{E}(X_{m(r)} | X_{o_1'(r)}) \right)^\top, \ldots, \left(X_{o(r)}, \widehat{E}(X_{m(r)} | X_{o_{q_r}'(r)}) \right)^\top \right\}^\top$ and $Y_{G(r)}$ with $(Y_{G(r)}^\top, \ldots, Y_{G(r)}^\top)^\top$ for final variable selection. (2) Averaged: The imputed value is the average of the above-mentioned stacked estimates. (3) Zero: Missing values are filled with zero, the mean of the observed values of each predictor. (4) SoftImpute: Missing values are imputed via the iterative soft-thresholded singular value decomposition method (Mazumder et al., 2010). (5) CC:

Only complete cases are used for variable selection if observations with complete modalities exist.

For each of the following setups, we repeat the simulation 100 times and set $B = 1000$. To evaluate and compare different methods, we use four measures, namely, ℓ_2 estimation error $\|\hat{\beta} - \beta^*\|_2$, denoted as ℓ_2-ER; false-positive rate (FPR); false-negative rate (FNR); and the elapsed time (in seconds) using R.

4.1 | Missing completely at random

We first generate the data without violation of our precondition about the missing mechanism, which is satisfied when the modalities of observations are missing completely at random (MCAR) as in the setting of Yu et al. (2020). For each model, the data are generated under two settings: (I) the dataset includes samples with complete observations and (II) the dataset does not includes samples with complete observations. For Setting (I), three modalities are considered with 50 predictors for each modality ($p = 150$). The dataset is composed of 200 samples with complete observations, 200 samples with observations from the first and second modalities, 200 samples with observations from the second and third modalities, and 200 samples with observations from the first and third modalities, as shown in the left panel of Figure 1. Thus, the total sample size is $n = 800$. For Setting (II), three modalities are considered with 20 predictors for each modality ($p = 60$). The dataset is composed of 500 samples with observations from the first and second modalities, 500 samples with observations from the second and third modalities, and 500 samples with observations from the first and third modalities. Thus, the total sample size is $n = 1500$. For each setting, we consider three models as follows:

Model 1 (linear regression): We first consider the following linear regression model:

$$Y = X\beta + \epsilon, \tag{10}$$

where $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n) \sim \mathcal{N}(0, 0.8^2 I_n)$, I_n is the n-dimensional identity matrix, and $(X_{i1}, \ldots, X_{ip})^\top \sim \mathcal{N}(0, \Sigma)$. The true regression coefficient vector is set by $\beta = (0.5\mathbf{1}_2^\top, 0_{48}^\top, 0.5\mathbf{1}_2^\top, 0_{48}^\top)^\top$. In Setting (I), Σ is a block diagonal matrix composed of 30 subblocks, where each subblock is a 5×5 square matrix with ones on the main diagonal and 0.5 elsewhere. In Setting (II), Σ is given by $\sigma_{jk} = 0.5^{|j-k|}$. The true coefficient vector is set by $\beta = (0.5\mathbf{1}_4^\top, 0_{16}^\top, 0.5\mathbf{1}_4^\top, 0_{16}^\top, 0.5\mathbf{1}_4^\top, 0_{16}^\top)^\top$.

Model 2 (logistic regression): We consider the logistic regression model (1) with parameter ξ_i for response Y_i given by

$$\xi = (\xi_1, \ldots, \xi_n)^\top = X\beta, \tag{11}$$

and $Y = (y_1, \ldots, y_n)^\top$ is sampled from the Bernoulli distribution with the success probability vector of $(e^{\xi_1}/(1 + e^{\xi_1}), \ldots, (e^{\xi_n}/(1 + e^{\xi_n}))^\top$. Other settings remain the same as those in Model 1.

Model 3 (Poisson regression): We generate Y from the Poisson distribution with mean vector $(e^{\xi_1}, \ldots, e^{\xi_n})^\top$, where $\xi = (\xi_1, \ldots, \xi_n)^\top$ is given in (11). For Setting (II), the true coefficient vector

TABLE 1 Performance comparison with missing completely at random data.

	Linear				Logistic				Poisson			
Method	ℓ_2-ER	FPR	FNR	Time	ℓ_2-ER	FPR	FNR	Time	ℓ_2-ER	FPR	FNR	Time
	Setting (I): $n = 800$, $p = 150$, with complete observations											
Stacked	0.576	0.392	0.000	31.66	0.985	0.433	0.000	35.18	0.708	0.504	0.000	35.89
(std)	0.065	0.132	0.000	1.37	0.117	0.108	0.000	0.48	0.075	0.130	0.000	0.61
Averaged	0.317	0.009	0.000	31.41	0.634	0.016	0.002	32.60	0.455	0.021	0.003	32.70
(std)	0.072	0.007	0.000	1.36	0.097	0.018	0.017	0.35	0.099	0.025	0.033	0.24
Zero	0.389	0.018	0.000	0.23	0.661	0.022	0.002	1.09	0.499	0.023	0.002	1.25
(std)	0.055	0.007	0.000	0.02	0.086	0.022	0.017	0.05	0.091	0.025	0.017	0.09
SoftImpute	0.381	0.016	0.000	653.1	0.656	0.022	0.002	642.5	0.490	0.023	0.002	636.9
(std)	0.059	0.007	0.000	16.04	0.094	0.025	0.017	8.99	0.093	0.024	0.017	23.05
CC	0.333	0.016	0.000	0.25	0.863	0.028	0.203	0.61	0.398	0.057	0.000	2.86
(std)	0.084	0.019	0.000	0.02	0.125	0.029	0.173	0.07	0.088	0.020	0.000	0.52
Proposed	0.240	0.011	0.000	38.83	0.503	0.020	0.007	39.10	0.360	0.020	0.005	39.11
(std)	0.065	0.009	0.000	5.23	0.130	0.022	0.033	5.08	0.101	0.043	0.029	5.32
	Setting (II): $n = 1500$, $p = 60$, without complete observations											
Stacked	0.432	0.098	0.000	5.52	0.854	0.150	0.001	7.46	0.631	0.358	0.008	8.32
(std)	0.053	0.056	0.000	0.07	0.069	0.076	0.008	0.23	0.101	0.174	0.038	0.43
Averaged	0.422	0.048	0.000	5.46	0.905	0.049	0.003	6.36	0.625	0.077	0.047	6.63
(std)	0.057	0.026	0.000	0.07	0.079	0.034	0.014	0.18	0.120	0.078	0.093	0.18
Zero	0.466	0.057	0.000	0.15	0.911	0.067	0.003	1.06	0.646	0.090	0.051	1.30
(std)	0.045	0.023	0.000	0.01	0.073	0.045	0.014	0.05	0.116	0.082	0.098	0.15
SoftImpute	0.462	0.059	0.000	181.1	0.917	0.064	0.002	181.2	0.639	0.092	0.048	183.6
(std)	0.048	0.025	0.000	4.09	0.067	0.038	0.012	5.38	0.119	0.084	0.095	5.00
Proposed	0.300	0.033	0.000	4.89	0.638	0.035	0.002	5.79	0.555	0.077	0.048	6.09
(std)	0.070	0.034	0.000	0.07	0.085	0.031	0.012	0.17	0.167	0.085	0.079	0.22

Note: The SDs of ℓ_2-ER, false-positive rate (FPR), false-negative rate (FNR), and time (in s) are listed in the corresponding second rows.

is changed to $\beta = (0.4\mathbf{1}_4^\top, 0_{16}^\top, 0.4\mathbf{1}_4^\top, 0_{16}^\top, 0.4\mathbf{1}_4^\top, 0_{16}^\top)^\top$. Other settings remain the same as those in Model 1.

Table 1 summarizes the comparison results for different methods. We obtain the following findings. First, the proposed method achieves the lowest estimation error ℓ_2-ER, FPR, regardless of the models considered. Second, the absence of complete observations does not significantly hinder the performance of the proposed method, since the proposed method can incorporate correlation information among covariates in every group. Third, the computing time of the proposed method is not the least because compared with other regression imputation-type methods that only regress missing covariates on observed ones, our approach regresses each covariate on all the other covariates. However, our method is much faster than the matrix completion-type method "SoftImpute." Moreover, we notice that the SDs are slightly higher for the proposed than

TABLE 2 Performance comparison with missing completely at random data under additional settings.

	Setting (III)				Setting (IV)			
Methods	ℓ_2-ER	FPR	FNR	Time	ℓ_2-ER	FPR	FNR	Time
Stacked	0.961	0.315	0.035	125.52	0.661	0.243	0.000	7.96
(std)	0.149	0.113	0.035	232.84	0.116	0.137	0.000	0.45
Averaged	0.484	0.019	0.007	125.20	0.497	0.027	0.011	7.95
(std)	0.147	0.030	0.034	232.56	0.150	0.025	0.060	0.45
Zero	0.490	0.017	0.007	0.23	0.499	0.027	0.000	0.10
(std)	0.145	0.017	0.034	0.17	0.131	0.026	0.000	0.01
SoftImpute	0.491	0.017	0.007	261.81	0.521	0.032	0.005	44.78
(std)	0.135	0.024	0.034	208.04	0.129	0.026	0.030	51.87
CC	0.491	0.023	0.000	0.17	0.506	0.051	0.022	0.11
(std)	0.139	0.023	0.000	0.13	0.165	0.042	0.095	0.01
Proposed	0.455	0.022	0.007	187.88	0.332	0.030	0.000	5.53
(std)	0.194	0.036	0.034	359.66	0.101	0.031	0.000	0.33

Note: The SDs of ℓ_2-ER, false-positive rate (FPR), false-negative rate (FNR), and time (in s) are listed in the corresponding second rows.

other methods in most cases because our algorithm involves a more complicated intermediate procedure that may induce additional variations in estimation.

To compare the proposed method with existing ones under a higher dimension with $p > n$, we consider Setting (III), which is similar to Setting (I) but with 60 predictors for each modality ($p = 180$). The dataset comprises 100 samples with complete observations and 20 from each of the three modality combinations, yielding a total sample size of $n = 160$. We consider Model 1 with the same setup, except that Σ is a block diagonal matrix composed of 36 subblocks. The true coefficient vector is set by $\beta = (0.5\mathbf{1}_2^\top, 0_{58}^\top, 0.5\mathbf{1}_2^\top, 0_{58}^\top, 0.5\mathbf{1}_2^\top, 0_{58}^\top)^\top$. Table 2 (left panel) summarizes the comparison results. Again, the proposed method has the lowest estimation error ℓ_2-ER and small FPR and FNR. Except for the naive "Zero" and "CC" approaches, all the methods take longer computing times, but our method is still more efficient than the matrix completion-type method "SoftImpute."

Furthermore, we consider a heavy-tailed case (Setting (IV)) to compare the performance of the proposed and existing methods when the normality assumption does not hold. Setting (IV) is similar to Setting (III) with the same configuration of the sample size but with 20 predictors for each modality ($p = 60$). We consider Model 1 with the same setup, except that $\epsilon_i \sim \sqrt{4/5}t(4)$, where $t(4)$ is the t distribution with the degree of freedom 4, and Σ is a block diagonal matrix composed of 12 subblocks. The true coefficient vector is set by $\beta = (0.5\mathbf{1}_2^\top, 0_{18}^\top, 0.5\mathbf{1}_2^\top, 0_{18}^\top, 0.5\mathbf{1}_2^\top, 0_{18}^\top)^\top$. Table 2 (right panel) presents the comparison results. We can draw similar conclusions; the proposed method has the lowest estimation error ℓ_2-ER, small values of FPR and FNR, and is much more efficient than "SoftImpute."

Finally, to examine whether the choice of B affects the estimation results, we re-analyze the datasets generated under Setting IV and Model 1 using $B = 500$. The results (not reported) show that the performance of our method is stable regardless of the values of B.

The R code can be downloaded at https://github.com/Yifan22Oct/BlockmissingGLM.

TABLE 3 Performance comparison with missing at and not at random (MAR and MNAR) data.

	MAR				MNAR			
Methods	ℓ_2**-ER**	**FPR**	**FNR**	**Time**	ℓ_2**-ER**	**FPR**	**FNR**	**Time**
Stacked	0.389	0.189	0.000	8.94	0.444	0.271	0.000	21.20
(std)	0.071	0.126	0.000	0.36	0.076	0.121	0.000	58.11
Averaged	0.314	0.014	0.000	8.86	0.330	0.023	0.000	21.09
(std)	0.067	0.015	0.000	0.35	0.054	0.017	0.000	58.11
Zero	0.312	0.017	0.000	0.14	0.328	0.018	0.000	0.16
(std)	0.057	0.014	0.000	0.01	0.064	0.014	0.000	0.01
SoftImpute	0.311	0.014	0.000	213.73	0.336	0.021	0.000	368.20
(std)	0.065	0.016	0.000	266.83	0.059	0.025	0.000	461.34
CC	0.377	0.039	0.000	0.11	0.522	0.063	0.000	0.12
(std)	0.091	0.029	0.000	0.01	0.081	0.049	0.000	0.03
Proposed	0.163	0.009	0.000	6.20	0.205	0.018	0.000	10.57
(std)	0.049	0.011	0.000	0.31	0.052	0.018	0.000	2.24

Note: The SDs of ℓ_2-ER, false-positive rate (FPR); false-negative rate (FNR), and time (in s) are listed in the corresponding second rows.

4.2 | Missing at/not at random

In this section, we compare the proposed and existing methods in terms of missing at and not at random (MAR and MNAR) data. Similar to Setting (I), three modalities are considered with 20 predictors for each modality and we add the fourth modality with 20 predictors $(X_{61},\ldots,X_{80})$. In Modality 4, covariates are fully observed across all groups. Similar to the setting 1 in Xue and Qu (2021), each sample is assigned to the complete-case group with probability proportional to $\exp(-a_i), i = 1,\ldots,n$, where $a_i = 10(X_{i,61} + \cdots + X_{i,80})$ in generating MAR data and $a_i = y_i$ in generating MNAR data. In addition, Σ is set to a block diagonal matrix composed of 16 subblocks, where each subblock is a 5×5 square matrix with ones on the main diagonal and 0.5 elsewhere. For simplicity, we only consider Model 1 (linear regression) and assign the regression coefficients as $\beta = (0.5\mathbf{1}_2^\top, 0_{18}^\top, 0.5\mathbf{1}_2^\top, 0_{18}^\top, 0.5\mathbf{1}_2^\top, 0_{18}^\top, 0.5\mathbf{1}_2^\top, 0_{18}^\top)^\top$. Other settings remain unchanged.

Table 3 provides a comparison of the estimation performance of the proposed and existing methods with MAR and MNAR data. The proposed method can reduce the selection bias caused by missingness and outperform the other methods even when the missing mechanism gets more complicated.

5 | REAL DATA ANALYSIS

In this section, we compare the proposed and existing methods in the ADNI data analysis. One of the main goals of the ADNI study is to recognize the biomarkers that can be used in clinical diagnosis and classification. We treat the binary diagnosis of whether a subject has AD or not as the response. The biomarkers to be selected are extracted from three sources in the ADNI-1 phase, namely, MRI, CSF, and proteomics datasets. The quantitative variables from MRI were processed

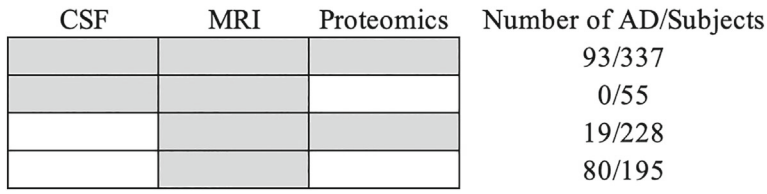

FIGURE 2 Pattern of block-wise missingness in the Alzheimer's Disease Neuroimaging Initiative dataset.

and measured by the UCSF team after cortical reconstruction and volumetric segmentation with the FreeSurfer image analysis site (Xiang et al., 2014). CSF samples were acquired at the University of Pennsylvania Medical Centre (Tzourio-Mazoyer et al., 2002). The proteomics biomarkers were delivered by the Biomarkers Consortium Plasma Proteomics Project. We screen out 100 features from MRI and proteomics source, respectively, through MV-SIS (Cui et al., 2015). For the CSF modality, five biomarkers, including amyloid β ($A\beta 42$), CSF total tau (t-tau), tau hyperphosphorylated at threonine 181 (p-tau), and two ratios ($A\beta 42/A\beta 40$ and $A\beta 42/A\beta 38$), were used. As shown in Figure 2, the dataset is composed of (1) 337 subjects with complete MRI, CSF, and proteomics features, of which 93 were AD, (2) 228 subjects with only MRI and proteomics features, of which 19 were AD, (3) 55 subjects with only MRI and CSF features, of which none was AD, and (4) 195 subjects with only MRI features, of which 80 were AD. Thus, $p = 205$, $n = 815$, and $R = 4$.

A test set of 100 samples is randomly drawn from the complete subjects 100 times to compare the performance of the proposed and existing methods, and the remaining subjects are used for training. For each data splitting, we fit the logistic regression model with ℓ_1 penalty to the training data. Table 4 reports the mean and SD of the classification errors and the mean model size for each method. Our proposed method exhibits the lowest classification error and selects relatively fewer biomarkers compared with the other methods.

Table S1 presents the biomarkers selected by each method based on the ADNI dataset. The proposed method determines 54 variables, including three CSF biomarkers, 16 MRI biomarkers, and 35 proteomics biomarkers. As shown in Table S1, our method tends to select variables with higher votes. At least four other methods also select most of the variables selected by our method. A few exceptions include t-tau, Aβ4240, ST90TA, Osteopontin, and Vascular Cell Adhesion Molecule-1. However, many previous studies have reported the associations between these biomarkers and AD progression. For example, Mattsson et al. (2016, 2018) found that patients with AD dementia had higher CSF t-tau than controls. Kwak et al. (2020) provided evidence to show that $A\beta 4240$ ratio reduction plays a critical role in AD therapy. Greene et al. (2010) revealed that the cortical thickness of the inferior parietal lobule in the right hemisphere ("ST90TA" is its average value) was significantly different among Cognitively Normal (NC), Mild Cognitive Impairment (MCI), and AD patients. Comi et al. (2010), Sun et al. (2013), and Carecchio and Comi (2011) asserted that the level of osteopontin correlates with cognitive declines because osteopontin is a molecule involved in macrophage recruitment and activation and implicated in neurodegeneration. Verbeek et al. (1994, 1995), and Huang et al. (2015) demonstrated the clinical significance of Vascular Cell Adhesion Molecule-1 to white matter disintegrity in Alzheimer's dementia based on the relationship between the receptor of T cell and the development of multiple sclerosis. The above evidence implies that the proposed method provides the analyst with a more targeted list of biomarkers sets, which can serve as a starting point for further study of AD pathogenesis.

TABLE 4 Performance comparison in the Alzheimer's Disease neuroimaging initiative data analysis.

Methods	FR	Size	Time
Stacked	0.1246	93.4500	26.9952
(std)	0.0254	17.9749	0.5135
Averaged	0.1199	51.9300	26.3098
(std)	0.0313	10.7217	0.4997
Zero	0.1210	52.6200	1.5653
(std)	0.0307	9.2188	0.1130
SoftImpute	0.1189	52.1300	221.3321
(std)	0.0287	10.8550	2.5717
CC	0.1258	46.4200	0.2649
(std)	0.0322	11.9892	0.0195
Proposed	0.1142	49.8500	37.2649
(std)	0.0322	9.1236	1.7647

Note: FR denotes the misclassification rate. Size denotes the mean model size. The SDs of FR, Size, and time (in s) are listed in the corresponding second rows.

6 | CONCLUSION

This paper proposes a single regression imputation method with multi-block missing data for high-dimensional GLMs. We develop a sparse inverse covariance-matrix estimation procedure by fully utilizing the structure of multi-block missing data from multi-modalities. The proposed estimator is positive semidefinite, and its ℓ_2 error bound is established via linear programming. Moreover, we impute the missing blocks through their means conditional on the observed blocks. Finally, a Lasso estimator of the coefficients in the GLM is obtained with the whole multi-modality data after imputation. Given that its imputation step is independent of the subsequent variable selection for GLM, the proposed method, which is markedly different from the existing approaches in dealing with multiblock missing data, can easily be extended to the other model context. Another superiority of the proposed method is that it mainly synthesizes the correlation information between covariates, and thus, can be implemented without complete observations. That is one of the main differences between the proposed and traditional regression imputation methods.

The derived theoretical results guarantee the effectiveness of the proposed method, and the simulation studies confirm that our approach is advantageous and more robust than several competing methods regardless of missingness mechanisms. An application to the ADNI study for identifying high-risk potential AD patients with biomarkers also demonstrates the utility and superiority of the proposed method. Finally, the proposed method conducts single imputation, but it can couple with the idea of multiple imputation (Harel & Zhou, 2007; Royston, 2004). Nevertheless, such advancement requires further investigation.

ACKNOWLEDGMENTS

We thank the Editor, Associate Editor, and two referees for their constructive comments and suggestions, which significantly helped improve our paper. This research was fully supported by

GRF grants 14301918 and 14302519 from the Research Grant Council of the Hong Kong Special Administrative Region and NSFC 11871263 from the National Natural Science Foundation of China.

ORCID

Xinyuan Song https://orcid.org/0000-0002-4877-3200

REFERENCES

Ando, R. K., Zhang, T., & Bartlett, P. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, *6*, 1817–1853.

Argyriou, A., Evgeniou, T., & Pontil, M. (2008). Convex multi-task feature learning. *Machine Learning*, *73*, 243–272.

Avella-Medina, M., Battey, H. S., Fan, J., & Li, Q. (2018). Robust estimation of high-dimensional covariance and precision matrices. *Biometrika*, *105*, 271–284.

Banerjee, O., El Ghaoui, L., & d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research*, *9*, 485–516.

Cai, T., Cai, T. T., & Zhang, A. (2016). Structured matrix completion with applications to genomic data integration. *Journal of the American Statistical Association*, *111*, 621–633.

Carecchio, M., & Comi, C. (2011). The role of osteopontin in neurodegenerative diseases. *Journal of Alzheimer's Disease*, *25*, 179–185.

Comi, C., Carecchio, M., Chiocchetti, A., Nicola, S., Galimberti, D., Fenoglio, C., Cappellano, G., Monaco, F., Scarpini, E., & Dianzani, U. (2010). Osteopontin is increased in the cerebrospinal fluid of patients with alzheimer's disease and its levels correlate with cognitive decline. *Journal of Alzheimer's Disease*, *19*, 1143–1148.

Cui, H., Li, R., & Zhong, W. (2015). Model-free feature screening for ultrahigh dimensional discriminant analysis. *Journal of the American Statistical Association*, *110*, 630–641.

Datta, A., & Zou, H. (2017). Cocolasso for high-dimensional error-in-variables regression. *The Annals of Statistics*, *45*, 2400–2426.

Fan, J., Feng, Y., & Wu, Y. (2009). Network exploration via the adaptive lasso and scad penalties. *The Annals of Applied Statistics*, *3*, 521–541.

Fan, Y., Demirkaya, E., & Lv, J. (2019). Nonuniformity of p-values can occur early in diverging dimensions. *Journal of Machine Learning Research*, *20*, 2849–2881.

Fan, Y., & Lv, J. (2013). Asymptotic equivalence of regularization methods in thresholded parameter space. *Journal of the American Statistical Association*, *108*, 1044–1061.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*, 432–441.

Greene, S. J., Killiany, R. J., & Alzheimer's Disease Neuroimaging Initiative. (2010). Subregions of the inferior parietal lobule are affected in the progression to alzheimer's disease. *Neurobiology of Aging*, *31*, 1304–1311.

Harel, O., & Zhou, X.-H. (2007). Multiple imputation: Review of theory, implementation and software. *Statistics in Medicine*, *26*, 3057–3077.

Huang, C.-W., Tsai, M.-H., Chen, N.-C., Chen, W.-H., Lu, Y.-T., Lui, C.-C., Chang, Y.-T., Chang, W.-N., Chang, A. Y., & Chang, C.-C. (2015). Clinical significance of circulating vascular cell adhesion molecule-1 to white matter disintegrity in alzheimer's dementia. *Thrombosis and Haemostasis*, *114*, 1230–1240.

Kwak, S. S., Washicosky, K. J., Brand, E., von Maydell, D., Aronson, J., Kim, S., Capen, D. E., Cetinbas, M., Sadreyev, R., Ning, S., Bylykbashi, E., Xia, W., Wagner, S. L., Choi, S. H., Tanzi, R. E., & Kim, D. Y. (2020). Amyloid-β42/40 ratio drives tau pathology in 3d human neural cell culture models of alzheimer's disease. *Nature Communications*, *11*, 1–14.

Lam, C., & Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of Statistics*, *37*, 4254–4278.

Lauritzen, S. L. (1996). *Graphical models* (Vol. *17*). Clarendon Press.

Li, Y., Yang, T., Zhou, J., & Ye, J. (2018). *Multi-task learning based survival analysis for predicting alzheimer's disease progression with multi-source block-wise missing data*. In *Proceedings of the 2018 SIAM international conference on data mining* (pp. 288–296). SIAM.

Liu, J., Ji, S., & Ye, J. (2012). Multi-task feature learning via efficient l2, 1-norm minimization. *arXiv preprint arXiv:1205.2631*.

Loh, P.-L., & Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, *40*, 1637–1664.

Mattsson, N., Smith, R., Strandberg, O., Palmqvist, S., Schöll, M., Insel, P. S., Hägerström, D., Ohlsson, T., Zetterberg, H., Blennow, K., Jögi, J., & Hansson, O. (2018). Comparing 18f-av-1451 with csf t-tau and p-tau for diagnosis of alzheimer disease. *Neurology*, *90*, 388–395.

Mattsson, N., Zetterberg, H., Janelidze, S., Insel, P. S., Andreasson, U., Stomrud, E., Palmqvist, S., Baker, D., Hehir, C. A. T., Jeromin, A., Hanlon, D., Song, L., Shaw, L. M., Trojanowski, J. Q., Weiner, M. W., Hansson, O., & Blennow, K. (2016). Plasma tau in alzheimer disease. *Neurology*, *87*, 1827–1835.

Mazumder, R., Hastie, T., & Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, *11*, 2287–2322.

Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, *34*, 1436–1462.

Royston, P. (2004). Multiple imputation of missing values. *The Stata Journal*, *4*, 227–241.

Städler, N., & Bühlmann, P. (2012). Missing values: Sparse inverse covariance estimation and an extension to sparse regression. *Statistics and Computing*, *22*, 219–235.

Sun, Y., Yin, X. S., Guo, H., Han, R. K., He, R. D., & Chi, L. J. (2013). Elevated osteopontin levels in mild cognitive impairment and alzheimer's disease. *Mediators of Inflammation*, *2013*, 615745–615749.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., & Joliot, M. (2002). Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *NeuroImage*, *15*, 273–289.

Verbeek, M. M., Otte-Höller, I., Westphal, J. R., Wesseling, P., Ruiter, D. J., & De Waal, R. (1994). Accumulation of intercellular adhesion molecule-1 in senile plaques in brain tissue of patients with alzheimer's disease. *The American Journal of Pathology*, *144*, 104–116.

Verbeek, M. M., Westphal, J. R., Ruiter, D. J., & De Waal, R. (1995). T lymphocyte adhesion to human brain pericytes is mediated via very late antigen-4/vascular cell adhesion molecule-1 interactions. *The Journal of Immunology*, *154*, 5876–5884.

Witten, D. M., & Tibshirani, R. (2009). Covariance-regularized regression and classification for high dimensional problems. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *71*, 615–636.

Xiang, S., Yuan, L., Fan, W., Wang, Y., Thompson, P. M., Ye, J., & Alzheimer's Disease Neuroimaging Initiative. (2014). Bi-level multi-source learning for heterogeneous block-wise missing data. *NeuroImage*, *102*, 192–206.

Xue, F., Ma, R., & Li, H. (2021). Semi-supervised statistical inference for high-dimensional linear regression with blockwise missing data. *arXiv preprint arXiv:2106.03344*.

Xue, F., & Qu, A. (2021). Integrating multisource block-wise missing data in model selection. *Journal of the American Statistical Association*, *116*, 1914–1927.

Yu, G., & Hou, S. (2022). Integrative nearest neighbor classifier for block-missing multi-modality data. *Statistical Methods in Medical Research*, *31*, 1242–1262.

Yu, G., Li, Q., Shen, D., & Liu, Y. (2020). Optimal sparse linear prediction for block-missing multi-modality data without imputation. *Journal of the American Statistical Association*, *115*, 1406–1419.

Yuan, L., Wang, Y., Thompson, P. M., Narayan, V. A., Ye, J., & Alzheimer's Disease Neuroimaging Initiative. (2012). Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. *NeuroImage*, *61*, 622–632.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, *11*, 2261–2286.

Yuan, M., & Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika*, *94*, 19–35.

Zhao, P., & Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, *7*, 2541–2563.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: He, Y., Feng, Y., & Song, X. (2023). Variable selection for high-dimensional generalized linear model with block-missing data. *Scandinavian Journal of Statistics*, 1–19. https://doi.org/10.1111/sjos.12632